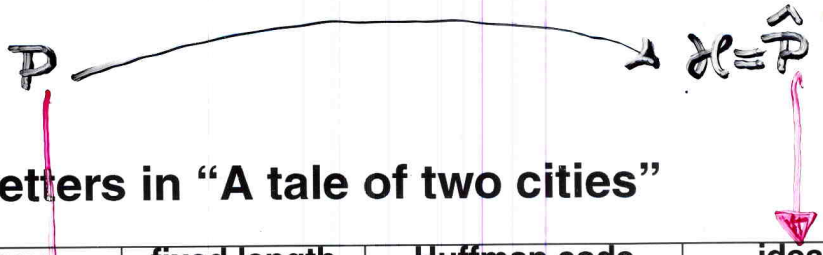
Glimpse of information theory



Coding letters in "A tale of two cities"

Letter	frequency		fixed length word	fixed length length	Huffman code word	Huffman code length	ideal length
a	47064	8.07 %	00000	5	1110	4	3.63
b	8140	1.40 %	00001	5	101111	6	6.16
c	13224	2.27 %	00010	5	01111	5	5.46
d	27485	4.71 %	00011	5	0110	4	4.41
e	72883	12.49 %	00100	5	000	3	3.00
f	13155	2.25 %	00101	5	111100	6	5.47
g	12120	2.08 %	00110	5	111101	6	5.59
h	38360	6.57 %	00111	5	1000	4	3.93
i	39786	6.82 %	01000	5	1010	4	3.87
j	622	0.11 %	01001	5	1111111110	10	9.87
k	4635	0.79 %	01010	5	11111110	8	6.98
l	21523	3.69 %	01011	5	10110	5	4.76
m	14923	2.56 %	01100	5	00111	5	5.29
n	41310	7.08 %	01101	5	1101	4	3.82
o	45118	7.73 %	01110	5	1100	4	3.69
p	9453	1.62 %	01111	5	101110	6	5.95
q	655	0.11 %	10000	5	1111111100	10	9.80
r	35956	6.16 %	10001	5	0010	4	4.02
s	36772	6.30 %	10010	5	1001	4	3.99
t	52396	8.98 %	10011	5	010	3	3.48
u	16218	2.78 %	10100	5	00110	5	5.17
v	5065	0.87 %	10101	5	1111110	7	6.85
w	13835	2.37 %	10110	5	01110	5	5.40
x	666	0.11 %	10111	5	1111111101	10	9.77
y	11849	2.03 %	11000	5	111110	6	5.62
z	213	0.04 %	11001	5	1111111111	10	11.42
total = 583.426		**100 %**	**mean = 5.00**		**mean = 4.19**		**H = 4.16**

Huffman $\approx$ *combinatorial entropy* (4.19 bits). Idealizing $\approx$ *entropy.* (4.16 bits). Theoretical units (nits rather than bits) corresponds to a change from base 2 to base e. Example also illustrates *redundancy.*

$\mathcal{H}$ permissable code (code length function) iff Kraft inequality holds, i.e.

$\sum 2^{-\mathcal{H}_i}=1$ or, in natural units, $\boxed{\sum e^{-\mathcal{H}_i}=1}$

NOTE: $P \sim \mathcal{H}$ amounts to $\mathcal{H}_i = \ln \frac{1}{P_i}$ or $P_i = e^{-\mathcal{H}_i}$